Comparison of Radiosonde and GCM Vertical Temperature Trend Profiles: Effects of Dataset Choice and Data Homogenization*

John R. Lanzante

NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, New Jersey

Melissa Free

NOAA/Air Resources Laboratory, Silver Spring, Maryland

(Manuscript received 16 October 2007, in final form 8 March 2008)

ABSTRACT

In comparisons of radiosonde vertical temperature trend profiles with comparable profiles derived from selected Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) general circulation models (GCMs) driven by major external forcings of the latter part of the twentieth century, model trends exhibit a positive bias relative to radiosonde trends in the majority of cases for both time periods examined (1960–99 and 1979–99). Homogeneity adjustments made in the Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC) and Hadley Centre Atmospheric Temperatures, version 2 (HadAT2), radiosonde datasets, which are applied by dataset developers to account for time-varying biases introduced by historical changes in instruments and measurement practices, reduce the relative bias in most cases. Although some differences were found between the two observed datasets, in general the observed trend profiles were more similar to one another than either was to the GCM profiles.

In the troposphere, adjustment has a greater impact on improving agreement of the shapes of the trend profiles than on improving agreement of the layer mean trends, whereas in the stratosphere the opposite is true. Agreement between the shapes of GCM and radiosonde trend profiles is generally better in the stratosphere than the troposphere, with more complexity to the profiles in the latter than the former. In the troposphere the tropics exhibit the poorest agreement between GCM and radiosonde trend profiles, but also the largest improvement in agreement resulting from homogeneity adjustment.

In the stratosphere, radiosonde trends indicate more cooling than GCMs. For the 1979–99 period, a disproportionate amount of this discrepancy arises several months after the eruption of Mount Pinatubo, at which time temperatures in the radiosonde time series cool abruptly by ~0.5 K compared to those derived from GCMs, and this difference persists to the end of the record.

1. Introduction

The potential utility of changes in vertical temperature structure of the atmosphere in the diagnosis of climate change has been recognized for several decades (Madden and Ramanathan 1980). While early studies identified such a climate change signal (Karoly 1987, 1989; Santer et al. 1996), recent work, using more advanced climate models and higher-quality radiosonde (Thorne et al. 2003) and satellite data (Santer et al. 2003), has yielded additional confirmation.

However, a closer examination of the problem revealed some systematic discrepancies. For example, a number of studies (Santer et al. 1996; Tett et al. 1996; Folland et al. 1998; Sexton et al. 2001; Tett et al. 2002; Hansen et al. 2002; Stott et al. 2006; Cordero and Forster 2006) found greater warming of the tropical upper troposphere in atmosphere–ocean general circulation models (GCMs) than in the observations. More broadly, some observations seemed to indicate that the surface had warmed more than the troposphere during the most recent 2–3 decades (National Research Council 2000), while climate models almost universally indi-

* Supplemental information related to this paper is available at the Journals Online Web site: http://dx.doi.org/10.1175/2008JCLI2287.s1.

Corresponding author address: John R. Lanzante, NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.
E-mail: john.lanzante@noaa.gov

DOI: 10.1175/2008JCLI2287.1

cated that the opposite should be occurring. Two separate committees of experts were convened to address this controversy (National Research Council 2000; Karl et al. 2006). The latter of these two groups concluded that previously undetected and/or uncorrected errors in the observed data were likely an important factor contributing to the discrepancy.

Ironically, reconciliation of the conflicting trend estimates was not the result of a narrower range of observational estimates, but instead an expanded range that better encompasses comparable estimates from climate models. Although estimates of large-scale, long-term temperature changes at the surface are reasonably well constrained, those from the highest-quality datasets available for the troposphere and the stratosphere still span a considerable range (Karl et al. 2006). The reason for the disagreements was tied primarily to data homogenization performed by different teams of analysts producing competing datasets from the same sets of raw observations. Homogenization is a crucial step in the production of datasets intended for use in assessing long-term climate change because of the potential corrupting effect of changes in instruments and recording practices that have occurred over time (Gaffen 1994). Homogenization is an attempt to eliminate the nonclimatic (i.e., artificial) component of change from the data. Because of the complexities and ambiguities involved in homogenization, competing teams of analysts can use different approaches, each of which seems scientifically defensible, yet create datasets whose trends are substantially different.

The intent of this paper is to address the following several fundamental questions: How much does homogenization influence the correspondence between observations and climate models? Are there any systematic effects or is the effect on this agreement random? Do different homogenized datasets behave differently? Are the effects similar for different time periods, latitude zones, and vertical layers? How do trends vary between different latitude zones and time periods? Because our primary interest is the comparison between observations and the models *as a collection*, we point out differences *between* models only in the most noteworthy instances. This study represents a follow-up to Lanzante (2007) in that it uses a similar methodological approach. However, it expands on that earlier work by employing both a larger, and more technically advanced, collection of climate models and observed datasets, and examines the stratosphere as well as the troposphere.

It is generally assumed that a suitably homogenized dataset renders more reliable estimates of climate change than the dataset consisting of the raw input observations. However, even if the homogenization process adds such value, there is no guarantee that improvement is found universally in all dimensions (*x*–*y*–*z*–*t*). For example, improvement may vary spatially and temporally, with little or no improvement, and perhaps degradation in some locales. Our purpose is not to validate observed datasets; indeed, the approach used is not suitable for such purposes. Instead, this work aims to explore the uncertainties that result from the homogenization process in the context of comparisons with temperature time series generated by climate models. Also, the purpose of this study is not that of formal detection and attribution of climate change, but rather is diagnostic in nature and thus serves as a complement to such work.

Observed and model data are introduced in section 2. Time series are used in section 3 to demonstrate the nature of the changes in temperature that have occurred. Radiosonde and GCM temperature trend profiles are compared and contrasted in section 4. Section 5 examines bivariate plots that summarize the effects of homogenization on the agreement between observed and model vertical trend profiles and assesses statistical significance. Summary and concluding remarks are given in section 6.

2. Observed and model data

a. Radiosonde observations

This work uses two state-of-the-art radiosonde datasets that have been extensively adjusted for homogeneity. Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC; Free et al. 2005) consists of station time series at 85 globally distributed locations for 13 vertical levels from the surface through the stratosphere (surface, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, and 30 hPa). The RATPAC dataset incorporates the Lanzante–Klein–Seidel (LKS; Lanzante et al. 2003a,b) homogeneity adjustments up through 1995 and uses the first differences technique thereafter (Free et al. 2004). We used the RATPAC-B version of this dataset. For some stratospheric applications we employ a reduced network of 47 RATPAC stations, referred to as RW, following Randel and Wu (2006) who eliminated some stations based on comparisons with independent satellite observations. The eliminated stations were deemed to have substantial artificial discontinuities in their stratospheric time series, even after homogeneity adjustment.

The other radiosonde dataset (Thorne et al. 2005) is the Hadley Centre Atmospheric Temperatures, version 2 (HadAT2), which consists of globally gridded time series for nine vertical levels above the surface (same as the RATPAC levels, except that it does not include the surface, 400-, 250-, and 70-hPa levels). Note that unlike RATPAC, HadAT2 does not contain surface data. Instead, following Karl et al. (2006), we use surface observations from the variance-adjusted Hadley Centre–Climatic Research Unit (CRU) surface temperature dataset (HadCRUT2v;[1] Brohan et al. 2006).

The radiosonde datasets are complementary in at least two important ways. First, RATPAC is based on a very thorough examination of a modest number of station time series. On the other hand, HadAT2 developers devoted less effort to each location, but instead utilized many more stations (nearly 700). Second, they employ fundamentally different techniques for homogenization. The LKS portion of RATPAC (up through 1997) uses multiple indicators in a labor-intensive expert committee approach, thereafter supplemented by the first-differences method. In large measure RATPAC does not depend on neighboring stations for identifying inhomogeneities or deriving subsequent adjustments; most of the information used is local to a given station. On the other hand, HadAT2 is based on comparisons with suitably chosen neighbors for both the identification and adjustment of its time series. As a result of these fundamental differences between RATPAC and HadAT2, use of both products serves as a check of the robustness of the results.

The radiosonde datasets are not completely independent because HadAT2 incorporates some adjusted station data from LKS. However, for several reasons we believe that they can still be considered largely independent products. First, only a modest fraction of the stations used to construct HadAT2 (57 out of nearly 700) are taken from LKS. Second, the homogenization methodology (local analysis versus neighbor checking), as discussed above, is fundamentally different. Third, through the HadAT2 iterative procedure any unique time histories of the LKS input will tend to be pushed toward a consensus of neighboring stations. Fourth, in the initial stages of constructing HadAT2, gross inconsistencies between some LKS stations and suitable non-LKS neighbors led to either the partial or total elimination of the suspect LKS records (Thorne et al. 2005).

RATPAC (additional information available online at http://www.ncdc.noaa.gov/oa/climate/ratpac/), HadAT2 (online at http://hadobs.metoffice.com/hadat/), and HadCRUT2v (online at http://www.cru.uea.ac.uk/cru/data/temperature) are freely available on the World Wide Web. These time series begin in 1958 and extend to near-present time via regular updates. Because a primary focus of this study is the effect of data homogenization, two versions of each radiosonde dataset were used—unadjusted (raw) and adjusted. The former consists of the data prior to any modifications aimed at homogenization. Although some of the raw input data to HadAT2 had been adjusted by LKS (see above), for simplicity we nevertheless refer to this input dataset as "unadjusted." (Note that this complete input set includes more stations than HadAT0.)

Although both radiosonde datasets have been derived with homogeneity as a primary goal, neither should be considered free from inhomogeneities. There is substantial evidence that trends derived from global collections of raw radiosonde temperature time series are systematically biased toward spurious cooling (Parker et al. 1997; Lanzante et al. 2003b; Sherwood et al. 2005; Karl et al. 2006). Even after homogenization these datasets are very likely to be afflicted by spurious cooling biases, perhaps substantial in nature (Sherwood et al. 2005; Randel and Wu 2006; Karl et al. 2006; McCarthy et al. 2008; Sherwood et al. 2008). It follows that differences in trends between adjusted and unadjusted versions of a radiosonde dataset are likely to underestimate the true effect of inhomogeneities in the data, perhaps by a considerable amount.

To examine temperature changes over large areas we focus on several broad latitude zones: Northern Hemisphere extratropics (NHX; 90°–30°N), tropics (30°N–30°S), and Southern Hemisphere extratropics (SHX; 30°–90°S), in addition to the entire earth (Globe; 90°N–90°S). Monthly values for each zone were formed by averaging all available station values into 10° latitude zones, and then into the mean for a given zone using cosine weighting by latitude.

The world-wide distribution of radiosonde stations is highly nonuniform (e.g., Fig. 1 of Lanzante et al. 2003a; Figs. 1 and 7 of Thorne et al. 2005), with many more stations over extratropical landmasses, particularly in the Northern Hemisphere, than over oceans. Also, fewer observations are available earlier in the period of record and at higher altitudes, such as in the stratosphere. As an example, for RATPAC there are roughly 35, 25, and 10 stations in the NHX, tropics, and SHX, respectively, going back to the early period of record in the troposphere (Table 2 of Lanzante et al. 2003b); comparable numbers for the satellite era (starting in 1979) are about 35, 30, and 15, respectively. Although we include the SHX for completeness, considerable caution is advised in interpreting results from this zone because it is sampled much more poorly than the others.

[1] Use of a more recent version of this dataset (HadCRUT3), which became available after the start of this project, does not materially affect our results. The change in surface temperature trend is at most 0.02 K decade^{-1} and usually less.

b. Climate model simulations

In support of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4), climate modeling groups worldwide performed coupled atmosphere–ocean climate model simulations of the twentieth century. All of the simulations used here include major anthropogenic (changes in greenhouse gases, ozone, the direct effects of sulfate aerosols) and natural (solar and volcanic) forcings. The details of these forcings, as well as additional forcings, and the number of ensemble runs vary between the models. The model data used here were obtained from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multimodel dataset. More detailed information on these integrations can be found online (http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php).

We have selected a subset of the available runs for use here, favoring those that used the largest number and type of natural and anthropogenic forcings. Two additional models that also included a wide range of forcings were not included because of resource constraints. Groups whose simulations were used include the Geophysical Fluid Dynamics Laboratory (GFDL), the Goddard Institute for Space Studies (GISS), the National Center for Atmospheric Research (NCAR), and the Met Office (UKMO). Two versions of the GFDL Coupled Model were used, CM2.0 and CM2.1, which differ in their dynamical core, cloud scheme, ocean viscosity, and land model. Similarly, two versions of the GISS Model, GISS-EH and GISS-ER, were used, differing only in the choice of ocean model. The version of the NCAR model used is the Parallel Climate Model (PCM), while for the UKMO the Hadley Centre Global Environmental Model version 1 (HadGEM1) was used. The number of ensemble members for each is as follows: CM2.0 (3), CM2.1 (3), PCM (4), GISS-ER (5), GISS-EH (5), and HadGEM (1).

Note that although the starting year for the GCM simulations is much earlier, we use model output starting in 1960, to match the time of the more widely available radiosonde observations. The ending year of our analyses (1999) coincides with the last year of most model simulations. Analyses focus on a longer period (1960–99), referred to as the radiosonde era, in addition to a shorter period (1979–99), referred to as the satellite era. Although satellite data are not employed in this study, the latter epoch is of interest because of the intense scrutiny that it has received in the literature.

3. Temperature changes from a time series perspective

Before examining vertical profiles of linear temperature trends, historical changes in temperature are examined via time series averaged globally over deep vertical layers in the troposphere (850–300 hPa) and stratosphere (100–50 hPa). In all of our comparisons, the GCM spatial (horizontal and vertical) and temporal sampling has been degraded to match that of the observations. In this section GCM time series are presented for the mean of the combined ensemble of 21 members, treating each member equally, along with the ensemble range, based on the minimum and maximum GCM values in a given month, and the adjusted versions of the observed datasets. The most obvious difference between the model and observed series is the presence of considerable short-term (multiyear) variability in the latter and an almost complete absence of such in the former. With a large sample (21) of coupled model realizations, each having its own unique (random) internal variability, the averaging process leads to cancellation, leaving mostly the forced variability, both natural and anthropogenic.

In the troposphere, adjusted versions of RATPAC and HadAT2 agree rather well at the global scale, and their shorter-term variability is dominated by ENSO (Fig. 1a). As expected, agreement for individual latitude zones (not shown), particularly on shorter time scales, is not as good. The observations almost always fall within the GCM ensemble spread. For longer time scales the observations track the GCM mean reasonably well, with both showing monotonic warming interrupted by major volcanic eruptions. The temperature drops dramatically for a year or so after each major eruption and then takes more than 5 yr to gradually recover (Santer et al. 2001).

Stratospheric series (Fig. 1b) show gross agreement between models and observations in that both indicate long-term cooling in the latter half of the record and dramatic short-term warming for a couple of years following each major volcanic eruption. While the GCM ensemble mean agrees reasonably well with the observations during the periods of volcanic warming, model spread is enhanced near the peak, with maximal warming typically about twice that of the ensemble mean.

During the satellite era the radiosonde cooling trend is greater than that in the GCMs (Fig. 1b). The elevation of temperatures for both observed datasets above that of the GCM mean just prior to the eruption of El Chichón is partly responsible. It is not clear whether long-term stratospheric cooling begins a few years sooner in the GCMs than the observations, or if inter-

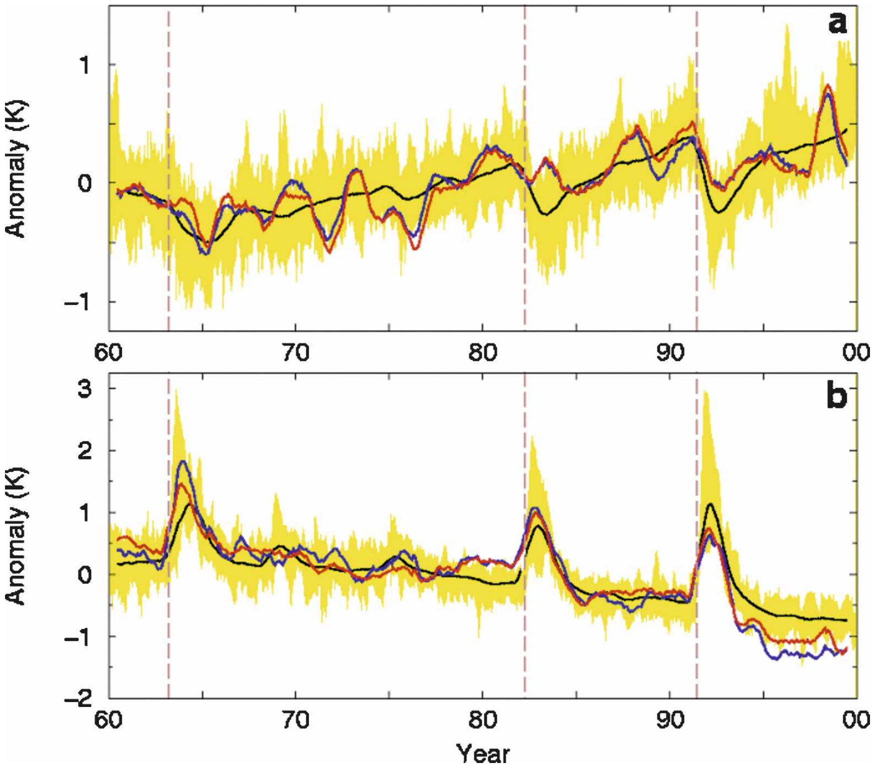


Fig. 1. (a) Globally averaged 850–300-hPa temperature time series for the GCM ensemble mean (black), adjusted RATPAC (blue), and adjusted HadAT2 (red). The ensemble spread (maximum to minimum) is shaded yellow. (b) Same as (a), but for 100–50-hPa temperature. All time series have been smoothed by applying a 12-point running average to monthly temperature anomalies (K) except for the ensemble minimum and maximum series, which are unsmoothed monthly values. To simplify this figure, the GCM monthly time series have been constructed by subsampling according to adjusted RATPAC only, both spatially and temporally. The GCM ensemble averages were computed utilizing the biweight mean (Lanzante 1996) to guard against the effects of outliers. Dashed vertical lines indicate times of major volcanic eruptions (Agung, March 1963; El Chichón, April 1982; and Mount Pinatubo, June 1991).

nal variability in the observations is responsible. Note that while the excess warmth of RATPAC and HadAT2 is quite consistent at the global scale (Fig. 1b), there are differences in the latitude zones (not shown). Although there is excellent agreement in the tropics, the excess warmth is found only in HadAT2 for NHX and only in RATPAC for SHX.

A larger contributor to the difference in trends develops several months after the Mount Pinatubo eruption. The magnitude of the disparity[2] is so large that the observed datasets lie near the GCM ensemble minimum (bottom of yellow fill in Fig. 1b); use of the RW network (not shown) yields a disparity close to that for HadAT2. The difference between the GCM and observed time series changes abruptly at this time and remains roughly constant thereafter, which implies a quick transition. For the individual latitude zones (not shown) this disparity is present in both NHX and the tropics, but not SHX. Because it seems unlikely that simultaneous homogeneity problems could occur and be undetected on such a large scale, this would imply that model and/or forcing deficiencies are a major contributing factor. On the other hand, if such homogeneity problems were to occur, the widespread presence of a simultaneous abrupt natural cooling signal could hamper detection and adjustment of the spurious signal in the observations. Cursory examination of satellite temperature records during this time period (Karl et al. 2006; WMO 2006) does not suggest any problems with

[2] The 50–100-hPa layer mean difference (model ensemble mean minus RATPAC mean) averaged over various time segments with starting and stopping points ranging from July 1991 to the end of 1993 is ~0.60–0.65 K.

the radiosonde observations, although uncertainties in the homogeneity of the satellite time series complicate any such inferences.

4. Vertical profiles of temperature trends

a. Troposphere

In this section various radiosonde and GCM vertical trend profiles are compared qualitatively as a prelude to section 5 where comparisons are made in a quantitative fashion. Linear trends are used here as a metric of climate change in the comparison of observed and model-generated temperatures. While some changes in temperature may not be strictly linear, this approximation is usually as valid as common alternates, such as step-like changes, that have been suggested in some instances (Seidel and Lanzante 2004). Trend estimates are based on the median of pairwise slopes (mps) method (Lanzante 1996) instead of ordinary least squares regression because, unlike mps, least squares has no resistance to outliers and is more sensitive to data near both ends of the record. As a result of the latter shortcoming, large-amplitude short-term climate anomalies (e.g., ENSO or volcanically related) that happen to fall near the ends of the record will have a disproportionate effect on least squares regressions. This consideration is particularly important when comparing observations with output from coupled climate models whose ENSO variability is not constrained to share the same phase as observations.

We have chosen to analyze the tropospheric and stratospheric portions of the temperature trend profiles separately for the following several reasons: 1) data quantity and quality are inferior in the stratosphere, 2) the formulations of the coupled climate models in the IPCC archive are more applicable to the troposphere, 3) the dominant physical mechanisms and forcings differ between the troposphere and stratosphere, and 4) trends in the stratosphere are an order of magnitude larger than in the troposphere. Items 1–3 suggest fundamental differences between the two regions and items 1 and 2 suggest more reliable results for the troposphere. Because of item 4, a combined analysis of both regions would yield results dominated by behavior in the stratosphere.

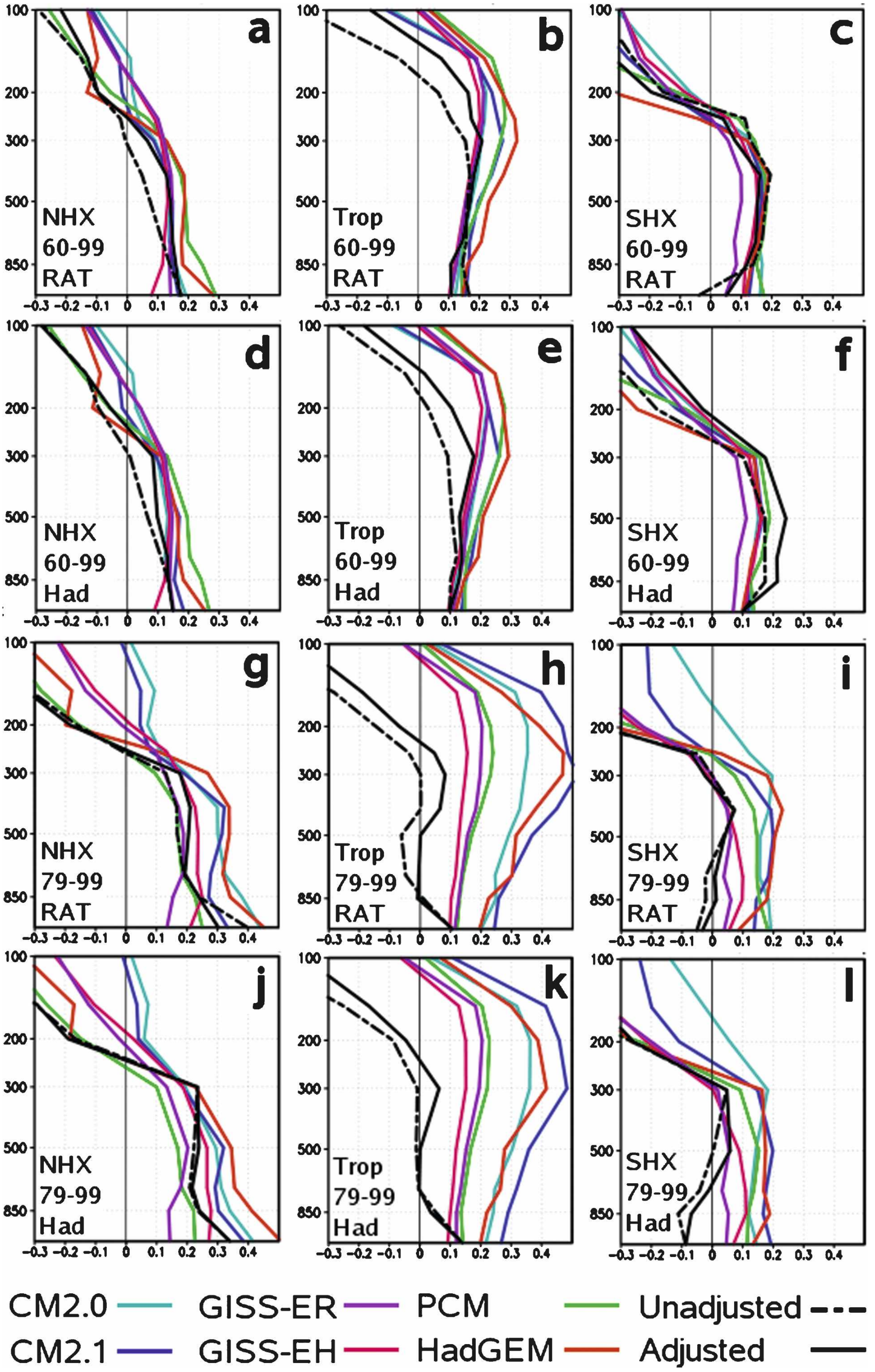
In the troposphere (Fig. 2), temperature trend profiles from the models, and to a somewhat lesser extent those from the observations, exhibit some distinctive differences between latitude zones. During the longer era, trends in the NHX (Figs. 2a,d) are largely uniform and positive in the main body of the troposphere, with a sharp decrease in the near-tropopause region leading to negative trends in the stratosphere. In the lower troposphere the models show a wide range of shapes, with some having much more warming at the surface than aloft, while others show similar trends at the surface and through most of the troposphere. Different boundary layer schemes as well as different methods used to derive surface temperature may contribute to the variety of model results.

The behavior in the SHX (Figs. 2c,f) is in some broad sense similar to that in the NHX (Figs. 2a,d), with a few exceptions. For models the exceptions are that the transition to cooling aloft occurs at a lower altitude and the surface and tropospheric trends are generally more consistent with each other in the SHX as compared to the NHX. For both the models and observations the exception is that the magnitude of the cooling aloft is larger in the SHX. These trend structural differences between the two extratropical regions may be related to differences in polar Antarctic and Arctic ozone depletion and associated forced cooling, and perhaps to differences in aerosol forcing as well. The behavior in the tropics is quite different with the transition to cooling aloft at a considerably greater altitude and an increase in warming upward from near the surface to a distinctive upper-tropospheric maximum. This upper amplification of the warming is readily explained by moist-adiabatic theory (Santer et al. 2005).

During the radiosonde era (Figs. 2a–f) the model

→

FIG. 2. Tropospheric temperature trend profiles for the (left) NHX, (middle) tropics, and (right) SHX over (a)–(f) 1960–99 and (g)–(l) 1979–99, based on (a)–(c), (g)–(i) RATPAC or (d)–(f), (j)–(l) HadAT2 radiosonde data. Trends are based on the mps technique (Lanzante 1996) and expressed in K decade^{-1} on the abscissa. The vertical height coordinate on the ordinate varies from the surface (bottom axis of each panel) upward in pressure coordinates (hPa). The black curves represent the trend profiles based on unadjusted (dashed) and adjusted (solid) radiosonde data. The colored curves are the corresponding ensemble mean trend profiles for the GCMs [CM2.0 (aqua), CM2.1 (blue), PCM (green), GISS ER (purple), GISS EH (magenta), HadGEM (red)] based on the same spatial and temporal sampling as the adjusted radiosonde data. The overall ensemble model spread corresponding to a subset of these figures is presented in the electronic supplement (available online at http://dx.doi.org/10.1175/2008JCLI2287.s1). Note that the surface values plotted for HadAT2 are taken from HADCRUT (Brohan et al. 2006) because the former does not have any. Because for HADCRUT there are no separate unadjusted and adjusted values, the values plotted should be considered adjusted. Note slight differences in model-derived profiles for the same latitude zone and time period based on analyses of RATPAC and HadAT2 resulting from subsampling in the GCMs spatially and temporally to match the adjusted observations.


a
NHX
60-99
RAT
b
Trop
60-99
RAT
c
SHX
60-99
RAT
d
NHX
60-99
Had
e
Trop
60-99
Had
f
SHX
60-99
Had
g
NHX
79-99
RAT
h
Trop
79-99
RAT
i
SHX
79-99
RAT
j
NHX
79-99
Had
k
Trop
79-99
Had
l
SHX
79-99
Had
100
200
300
500
850
−0.3 −0.2 −0.1 0 0.1 0.2 0.3 0.4
CM2.0
CM2.1
GISS-ER
GISS-EH
PCM
HadGEM
Unadjusted
Adjusted

profiles show reasonable consistency among themselves in the free atmosphere. The radiosonde profiles, while exhibiting the gross differences between latitude zones noted above for the models, often lie outside of the envelope of intermodel spread of the ensemble means. The discrepancy in the tropical upper troposphere is particularly noteworthy. Adjustment almost always enhances tropospheric warming and improves radiosonde–model agreement, with exceptions in the SHX where radiosonde sampling is more limited.

Trend profiles from the two radiosonde datasets show considerable similarity, and in general show more similarity with one another than with the models. The change in trend resulting from adjustment is comparable in magnitude to the intermodel spread of the ensemble means for much of the troposphere in the NHX and tropics. This complements the finding of Lanzante (2007) that the effects of adjustment are sometimes comparable to the effects of a major model forcing, further highlighting the importance of homogeneity adjustment in evaluating radiosonde temperature trends.

In comparing trend profiles from the two time periods, the most obvious difference is the larger intermodel spread during the satellite as compared to the radiosonde era. This is consistent with the larger statistical uncertainty caused by the confounding effects of short-term climate noise resulting from internal variability (Stott and Tett 1998), for example, ENSO. The large spread is especially prominent for the tropics (Figs. 2h,k) where two of the three models with the largest tropospheric warming have unrealistically large ENSO amplitudes (CM2.0, and especially CM2.1; see Wittenberg et al. 2006), while the other (HadGEM) has only one ensemble member. Another difference between epochs is the greater stratospheric cooling and associated lowering of the altitude of transition between warming and cooling for the satellite era; this is not surprising because most of the stratospheric ozone depletion has occurred during the satellite era. The observed temperature profiles in the SHX for the satellite era are unusual, with cooling at the surface but warming in the free troposphere; none of the model trend profiles for this or any other region or time period shows this pattern. Sampling may be one explanation because the surface trend from the more spatially complete HadCRUT data is slightly positive (not shown).

Overall, observation–model agreement is poorer during the satellite era, both in terms of the overall magnitude of tropospheric warming as well as the shapes of the profiles. Nevertheless, results from both eras are in general agreement that 1) adjustment makes trends more positive and 2) adjustment enhances the agreement between models and observations.

b. Stratosphere

Long-term temperature change in the stratosphere (Fig. 3) is characterized by cooling at almost all levels for almost all cases, with trend magnitudes several times larger than in the troposphere. Although cooling generally increases with height in the lower stratosphere, in going from 50 to 30 hPa the cooling often decreases. However, this tendency varies considerably from case to case, and is most apparent in the GFDL simulations (CM2.0 and CM2.1).

While observation–model agreement is reasonably good overall with regard to profile shape, with regard to magnitude the observations show consistently and often considerably more cooling than the models. As for the troposphere, the intermodel spread is noticeably larger during the shorter period. It is worth noting that the PCM stands out as having the poorest agreement with both the observations and the other models. In the two extratropical zones the shapes of the PCM profiles typically bear little resemblance to all of the other profiles. In the tropics, although the shape is much more similar to the others, the PCM has noticeably less cooling that the other models, especially resulting from behavior at the highest altitudes. Other work, which will be reported in a separate manuscript, shows that the PCM also differs strongly from the other models in its simulation of the stratospheric temperature response to volcanic eruptions. The GFDL models and RATPAC indicate more cooling at 50 hPa in the tropics than in the extratropics in contrast to earlier work that showed most models having more cooling in the Southern Hemisphere extratropics than elsewhere (Shine et al. 2003).

Comparison of models and observations in the stratosphere yields broad conclusions similar to that for the troposphere. The model minus observed trend difference is almost always positive, and this difference is larger during the satellite than the radiosonde era. The magnitude of this difference is roughly comparable to the model trends. The effect of adjustment is usually to decrease the estimated observed stratospheric cooling and the model–observation difference. The effect of adjustment is largest in the tropics, especially for the satellite era, consistent with the notion that tropical radiosonde temperature data are particularly subject to time-varying measurement biases (Lanzante et al. 2003b; Sherwood et al. 2005; Randel and Wu 2006). Finally, trends from the two observed radiosonde datasets are found to be more similar to each other than they are to comparable trends derived from models. Observed trends often lie outside the envelope of spread of the model ensemble means.

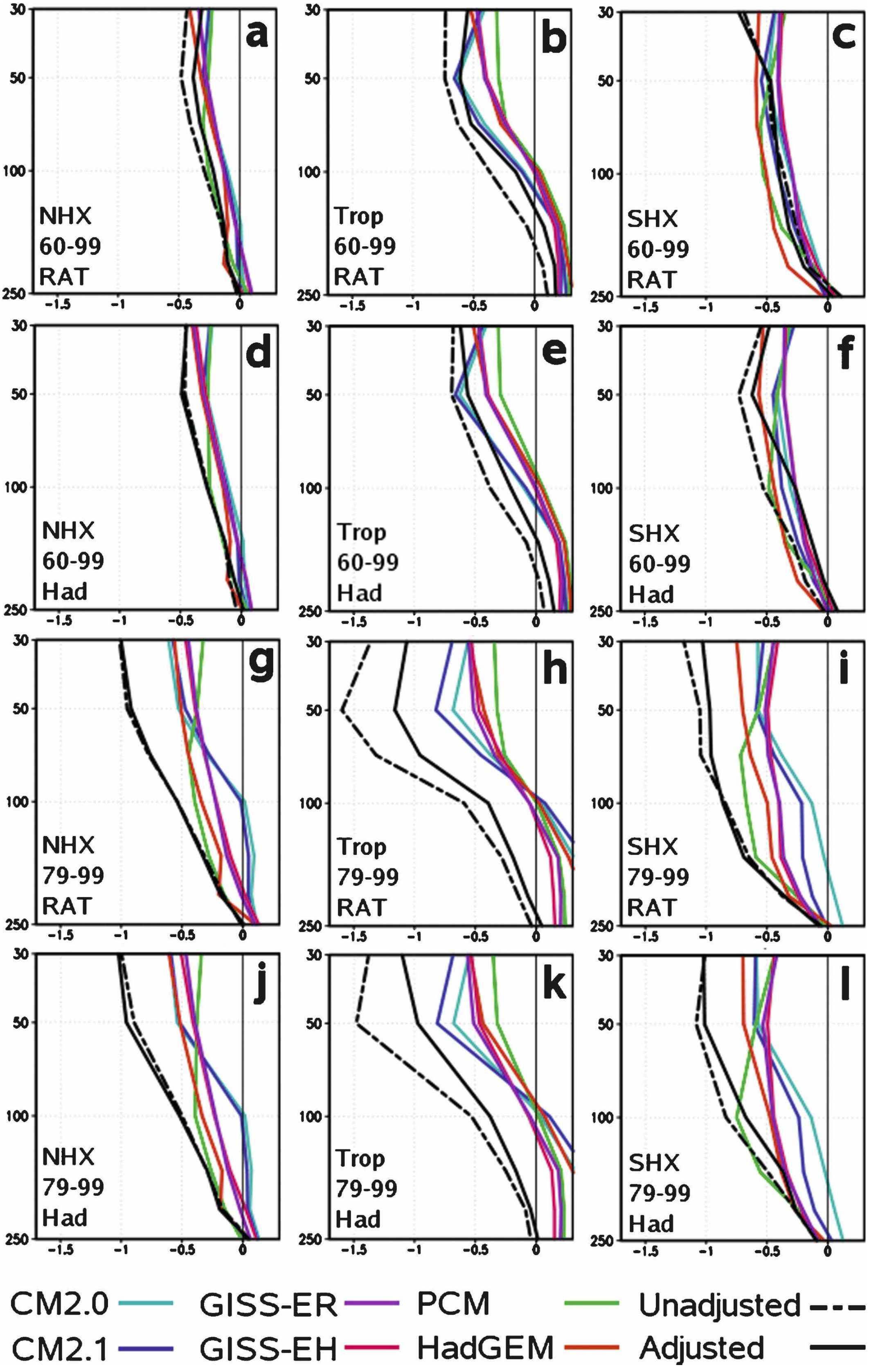


FIG. 3. Same as Fig. 2, but for the stratosphere with a correspondingly different range along the ordinate.

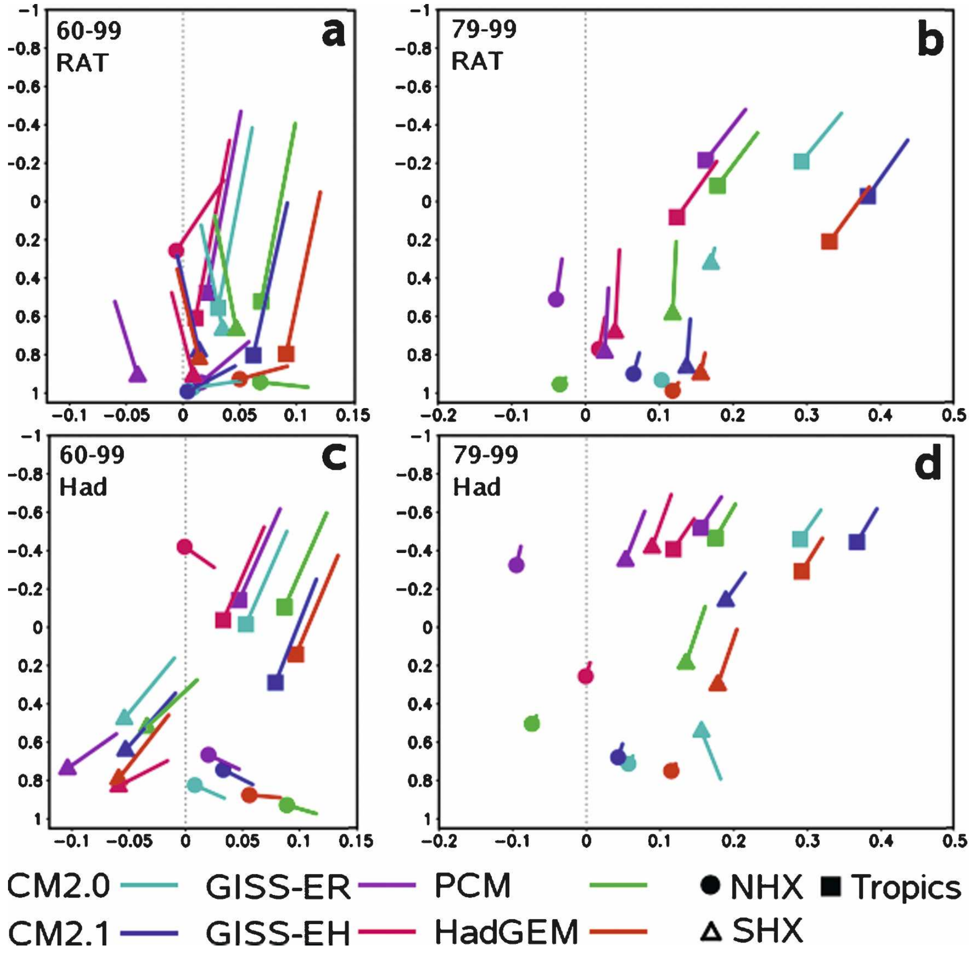


FIG. 4. Bivariate metrics of comparison of temperature trend profiles between GCMs and observations for the troposphere (surface–250 hPa for NHX and SHX; surface–150 hPa for the tropics). As indicated in the legend, results are color coded for the GCMs, and designated by symbols for latitude zones. Each point corresponds to the comparison between one observed dataset and the ensemble mean for one GCM. Metrics for each ensemble member corresponding to a subset of these figures are presented in the electronic supplement (available online at http://dx.doi.org/10.1175/2008JCLI2287.s1). The horizontal coordinate is the difference, GCM minus observed, of the layer-averaged temperature trend while the vertical coordinate is the correlation between the GCM and observed profile over the layer. Each line connects the metrics based on adjusted (symbol end) and unadjusted (opposite end) observed data. Perfect model–observation agreement would be indicated by a coordinate of (0, 1). Note that the ordinate increases downward. (a), (b) RATPAC and (c), (d) HadAT2 radiosonde data based on trends for (a), (c) 1960–99 and (b), (d) 1979–99.

5. Quantitative assessment of temperature trend profiles

a. Bivariate plots

As a complement to the qualitative evaluations made in section 4, a quantitative approach is used here based on a slight modification of the scheme developed by Lanzante (2007). A bivariate metric of agreement between the trend profiles from one model and one observed dataset is plotted as an *x*–*y* pair, as, for example, in Fig. 4. The abscissa is the difference in the layer mean trends, the model minus observed data, and as such it represents a measure of relative bias.[3] The ordinate, which is the correlation between the model and observed trend profiles in the layer, represents a measure of agreement in the shapes of the profiles. Because

[3] Note that the order of differencing in defining the bias is arbitrary and is not meant to imply that the observed trends are "correct." Any differences could be due to observational errors, model errors, or a combination of both.

they are based on summary statistics, examination of various bivariate plots helps yield insights not as readily apparent from visual examination of the corresponding trend profiles.

The results summarized in Fig. 4a for RATPAC radiosonde era tropospheric trends correspond to the profiles shown in Figs. 2a–c. The fact that almost all lines have positive x coordinates indicates that layer mean trends for the models are consistently greater than those derived from RATPAC. For NHX the correlations are quite high, indicating excellent shape agreement. Data adjustment has the effect of decreasing the relative bias, which nearly vanishes for some models. Results for the tropics show that while adjustment has a comparable effect in uniformly reducing relative bias, it has a much more dramatic effect in increasing shape agreement. As seen in Fig. 2b, shape agreement is poor prior to adjustment but becomes good after adjustment as warming is increased in the upper troposphere, moving the maximum upward. For the SHX the effect of adjustment is primarily to increase agreement in shape, yielding excellent overall agreement, as seen in Fig. 2c as well. Overall, aside from the tendency for somewhat more warming in the models than the observations, adjusted RATPAC trends yield reasonably good agreement with the models for this time period.

As for RATPAC, adjustment of HadAT2 (Fig. 4c) reduces the positive relative bias for NHX and the tropics, but in contrast it increases a negative relative bias in the SHX. While adjustment enhances shape agreement for the SHX and tropics, it slightly degrades shape agreement for the NHX. In the tropics, trends from adjusted HadAT2 do not agree with the models as well as RATPAC, as seen in the trend profiles (Fig. 2e versus Fig. 2b), because the upper-tropospheric warming and maximum are not as pronounced. The poorer relative bias agreement in the SHX for HadAT2 as compared to RATPAC and its degradation with adjustment are also evident in the corresponding trend profiles (Fig. 2f versus Fig. 2c). Overall, for radiosonde era tropospheric trend profiles the agreement with models is better for RATPAC than HadAT2.

For the satellite era in the troposphere (Figs. 4b,d) some aspects of the trend comparisons resemble those of the radiosonde era. Positive relative biases dominate with better agreement for RATPAC than HadAT2 and better agreement after adjustment. In addition, agreement in the tropics is poorer than in the NHX and SHX. However, in contrast to the radiosonde era, the agreement is often considerably poorer and the intermodel spread is much greater; both of these tendencies are exhibited clearly in the trend profiles as well (Figs. 2g–l), especially for the tropics.

Stratospheric bivariate plots for RATPAC and HadAT (Figs. 5a–d) have a distinctly different character than those for the troposphere. Except for a few instances, adjustment has very little effect on shape agreement. This may be due in part to the fact that the stratospheric trend profiles have less complexity than those for the troposphere. In NHX and the tropics shape agreement is excellent, with correlations exceeding 0.9 in most cases. The dominant effect of adjustment is typically to reduce the positive relative bias, the magnitude of which can be several times larger than that in the troposphere. However, even after adjustment there is often a big relative bias, especially during the satellite era (Figs. 3g–l) when observed cooling can be approximately twice as large as in the models.

For the stratosphere, observation–model agreement is quite different in the SHX compared to the other two zones because shape agreement is often much poorer and the intermodel spread in shape agreement is often quite large. Sampling variability may play a role in the unusual behavior in the SHX because the number of stations is far fewer than for the other zones. This deficiency is magnified by the fact that there are fewer observations at higher altitudes because of premature bursting of radiosonde balloons that is more likely at the low temperatures and pressures of the stratosphere. Another factor that may play a role in the odd behavior in the SHX is the large radiative forcing resulting from stratospheric ozone depletion. Uncertainties either in this forcing or in the feedbacks involved may act to magnify the differences between different model simulations. Finally, as noted in section 4b, PCM is an outlier in the bivariate plots, especially for the SHX, but also for the NHX.

The bottom two panels in Figs. 5e,f involve trends from a reduced RATPAC network (RW) based on eliminating some stations whose adjusted time series appear to be corrupted by artificial inhomogeneities in the stratosphere (Randel and Wu 2006). In these plots the starting point of each line is based on the full RATPAC network, while the symbol end corresponds to the reduced network, and as such these figures show the effects of eliminating the suspect stations. In the interest of brevity, the trend profiles for the reduced network are not shown.

During the radiosonde era, for NHX and the tropics, reducing the network has only a minor effect, whereas during the satellite era there is a considerable reduction in relative bias. This is not surprising because Randel

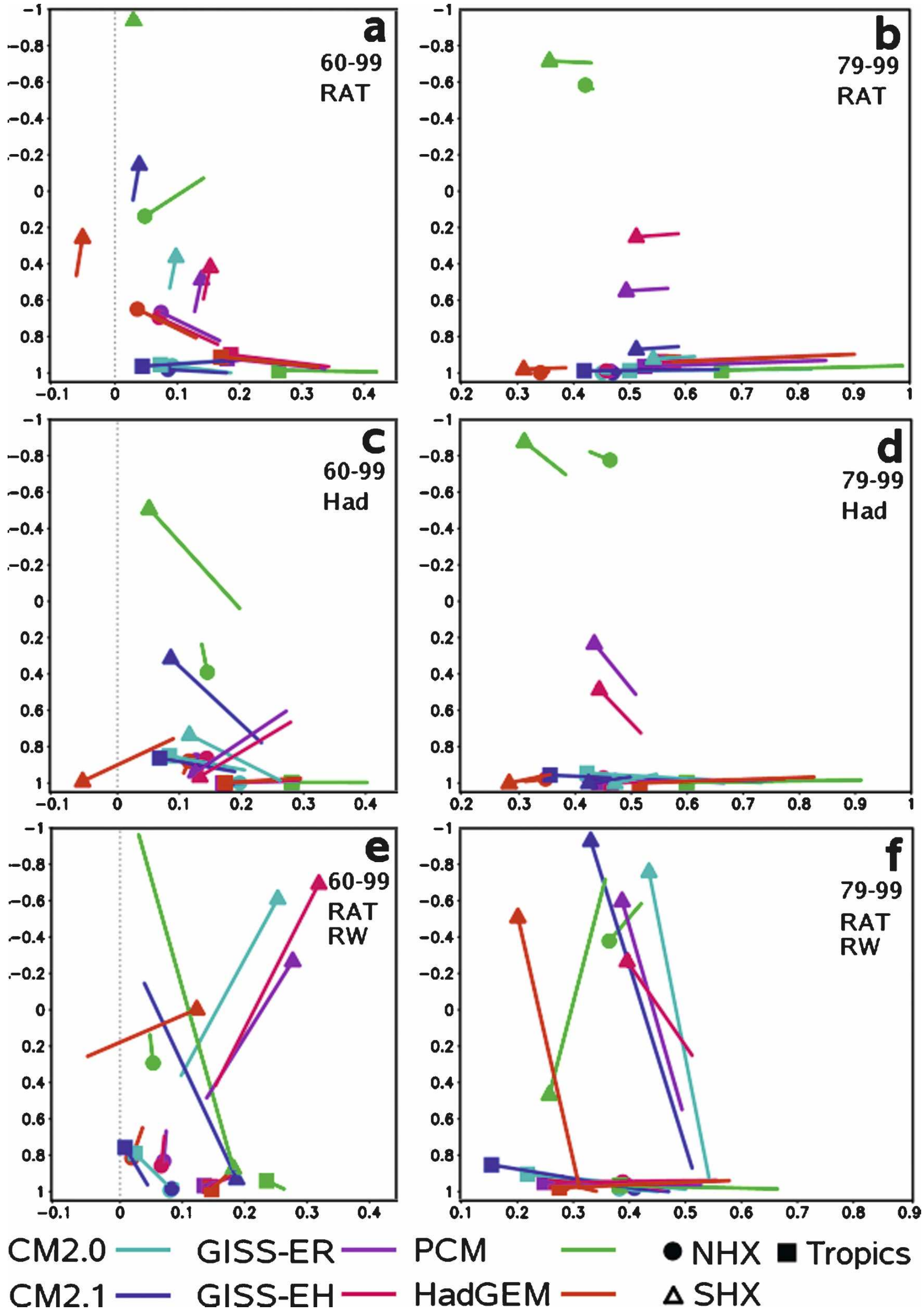


FIG. 5. Same as Fig. 4, but for the stratosphere (100–30 hPa) based on trends for (left) 1960–99 and (right) 1979–99. Furthermore, for (e) and (f) only, the symbol end of each line corresponds to adjusted data for the RW, whereas the opposite end is based on adjusted data for the full RATPAC network.

and Wu (2006) eliminated stations whose behavior was suspect during the satellite era. Furthermore, the nature of the problems indicated an overall spurious cooling bias prior to elimination. The behavior in the SHX is quite unexpected because network reduction greatly degrades shape agreement for most models, but for the PCM it greatly improves it. Network reduction results in observed profiles that have mostly decreasing cooling with altitude above 100 hPa (not shown). Determining whether enhanced agreement with the PCM occurs

Table 1. Summary of results from Tables A1–A6, excluding those based on the RW dataset. Each number is the percentage of cases deemed significant at the 5% level. Tests involve the components of the bivariate metric (difference, correlation). The first (second) row involves testing whether the GCM and observed layer mean trends are different (trend profiles are correlated). Row 3 (4) reports whether homogeneity adjustment reduces the GCM–observation difference (increases the GCM–observation correlation). Row 5 (6) indicates whether one dataset has better agreement with the GCMs with regard to the difference (correlation) metric; R indicates RATPAC is better and H indicates HadAT2 is better.

Tested	Troposphere	Stratosphere
(GCM–observation) difference	67	100
(GCM versus observation) correlation	46	75
Adjustment reduces difference	33	58
Adjustment increases correlation	75	17
(RATPAC versus HadAT2) difference	R: 17 H: 0	R: 17 H: 17
(RATPAC versus HadAT2) correlation	R: 100 H: 0	R: 0 H: 33

for the wrong reasons or whether the PCM should be considered a superior performer in the stratosphere is beyond the scope of this work.

b. *Statistical significance*

Statistical significance testing has been applied to a number of hypotheses involving the bivariate measures of observation–model agreement. The nature of these tests and detailed results are given in the appendix. A brief overview is given here based on summary statistics given in Table 1. A cell in Table 1 represents an aggregate measure derived from more detailed results. Each number in Table 1 is the percentage of individual tests in the aggregate pool found significant at the 5% level or better. A particular pool includes aggregation over several of the following categories: trend time period (1960–99/1979–99), latitude zone (NHX/tropics/SHX), dataset (RATPAC/HadAT2), and homogeneity treatment (unadjusted/adjusted).

The first row in Table 1 deals with testing whether the GCM and radiosonde layer mean trends are different. The results indicate that such differences are significant for all cases in the aggregate pool in the stratosphere and most cases in the troposphere. The second row, involving comparison of the shapes of the GCM and radiosonde trend profiles via the correlation metric, indicates that they are significantly correlated for most cases in the stratosphere and almost half of the time for the troposphere. Results from the third and fourth rows indicate that homogeneity adjustment often improves the agreement between observed and GCM trend profiles, with larger improvement in the troposphere regarding the shape of the profiles and in the stratosphere regarding the layer mean trend. The final two rows focus on whether one of the two radiosonde datasets has a significantly better agreement with the GCMs. Regarding the difference in layer mean trend (row 5) neither radiosonde dataset exhibits much of an advantage. However, regarding agreement in the profile shape between the GCMs and observations, RATPAC is always better in the troposphere while HadAT2 is sometimes better in the stratosphere.

Broadly speaking, several conclusions can be drawn from Table 1. First, there are considerable differences in layer mean trends between GCMs and radiosondes (row 1), especially in the stratosphere. Homogeneity adjustment has a major impact in reducing these differences, particularly in the stratosphere (row 3). There is a moderate degree of agreement in the shapes of the GCM and radiosonde trend profiles (row 2). Shape agreement is poorer in the troposphere (row 2) perhaps because of the greater complexity of the profiles there. Adjustment improves shape agreement slightly in the stratosphere, but considerably in the troposphere (row 4). Regarding agreement between GCM and radiosonde layer mean trends, the two radiosonde datasets do not differ greatly (row 5). However, regarding shape, HadAT2 agrees slightly better with GCMs in the stratosphere while RATPAC has an overwhelmingly better agreement in the troposphere.

Regarding the differences between RATPAC and HadAT2 in their agreement with GCMs, one can only speculate as to the cause. First, we should caution that better agreement with GCMs does not necessarily imply a more realistic product. The more-detailed and labor-intensive procedure used to produce LKS (which constitutes a major part of RATPAC) has the potential to better delineate details in vertical structure, perhaps explaining the tropospheric result in row 6. On the other hand, in the stratosphere, where frequent balloon bursts at the highest altitudes can greatly reduce the quantity of data, the HadAT2 team may benefit from their use of a much more extensive network of stations, as well as the "neighbor information" that their method utilizes.

6. Conclusions and discussion

Temperature trend profiles derived from observed radiosonde data have been compared with comparable trend profiles derived from GCMs driven by major natural and anthropogenic forcings from the latter part of the twentieth century. These comparisons have been made visually and by calculating the difference in layer mean trends and the correlation between the trend profiles. The principal findings are as follows:

1) Homogeneity adjustment of radiosonde temperature time series almost always improves agreement between observations and model trend profiles. This improvement is almost always statistically significant in the tropical troposphere and sometimes significant elsewhere. The change in trend resulting from adjustment is often as large as the difference in ensemble means between the different models.
2) Agreement in the trend profiles between the models and observations is moderately better for RATPAC than HadAT2 in the troposphere, but slightly better for HadAT2 in the stratosphere. Agreement between RATPAC and HadAT2 is usually much better than the agreement between either dataset and the models.
3) Model trends exhibit a consistent, mostly significant positive bias relative to the observations, which are generally larger, (a) in the stratosphere than the troposphere, (b) during 1979–99 rather than 1960–99, (c) for unadjusted rather than adjusted observations, and (d) in the tropics rater than the extratropics.
4) In the troposphere, the tropics have both the poorest agreement between models and observations and the largest improvement via homogeneity adjustment. Adjustment increases layer mean warming and moves the level of maximum warming upward. Nevertheless, even after adjustment the observed and model trend profiles show considerable disagreement.
5) During the satellite era stratospheric cooling is significantly greater in the observations than in the models. A considerable amount of this is due to a difference that develops several months after the Mount Pinatubo eruption and persists through the end of the record, such that observations are ~0.5 K cooler than models.

By itself, the fact that homogeneity adjustment consistently and overwhelmingly enhances agreement between observation and model trend profiles does not imply that the adjusted data are closer to reality. However, this finding should be viewed in light of the mounting evidence that (a) raw radiosonde temperature time series are corrupted by extensive artificial inhomogeneities that impart a systematic cooling bias (Parker et al. 1997; Lanzante et al. 2003b; Sherwood et al. 2005; Karl et al. 2006), and (b) the impact of the biases on temperature trends varies in magnitude with altitude (Lanzante et al. 2003b; Sherwood et al. 2005; Karl et al. 2006). Thus, successful elimination of some component of these biases will alter both the layer mean trend and shape of the profiles in such a way to produce more realism. Furthermore, the two adjusted radiosonde datasets used here yield consistently better model agreement than their unadjusted counterparts, were created via fundamentally different input data and methodology, and are both completely independent of GCMs. Taken together, these facts suggest that a substantial amount of the improved agreement is likely due to a more realistic representation of observed temperature trends resulting from homogeneity adjustment. Similarly, remaining differences in trends between the models and adjusted observations could be explained at least in part by substantial unresolved inhomogeneities in radiosonde data, which likely contaminate the adjusted data (Sherwood et al. 2005; Randel and Wu 2006; Karl et al. 2006; McCarthy et al. 2008; Sherwood et al. 2008). This is consistent with the conclusion by McCarthy et al. (2008) that uncertainties in adjustment methodology are sufficiently large to explain tropical trend discrepancies between HadAT2 and theoretical expectations.

The above statements are not meant to imply that the model simulations are free from important errors or systematic biases. For example, the abrupt separation of observed and model temperatures beginning several months after the Mount Pinatubo eruption appears to be responsible for a good deal of the discrepancy in stratospheric trends between models and radiosonde observations over the satellite era. There is no evidence of widespread drops in radiosonde temperatures associated with artificial inhomogeneities during the narrow timeframe identified. Furthermore, enhanced ozone loss and increased stratospheric water vapor resulting from volcanic aerosols immediately after eruptions (Solomon et al. 1996; Joshi and Shine 2003) are not included in the forcings for these model simulations, and might play a role (Ramaswamy et al. 2006). Taken together, this evidence, which suggests possible shortcomings in model simulations (i.e., model formulations and/or forcings) that could explain some of the trend discrepancy, provides motivation for future research.

Acknowledgments. We acknowledge the modeling groups for making their simulations available for analysis, the Program for Climate Model Diagnosis and In-

Table A1. Significance level (%) for metrics comparing the vertical temperature trend profile from one radiosonde dataset (RATPAC or HadAT2) with the corresponding profiles from the combined GCM ensemble pooled from all models. Trend profiles are for the troposphere (surface–250 hPa for NHX and SHX; surface–150 hPa in the tropics) for (left) 1960–99 and (right) 1979–99. Two metrics are assessed: (a) relative bias (diff) as measured by the difference in layer mean trend, GCM minus observation, and (b) shape (corr) as measured by the correlation between an observed and GCM trend profile. Rows correspond to the latitude zones over which the profiles were derived. The two numbers in each cell correspond to (left) unadjusted and (right) adjusted versions of the observed dataset. The algebraic sign for the diff metric indicates whether the GCM trends are larger (positive) or smaller (negative) than those from observations, and for the corr metric indicates whether the correlation is greater than or less than zero. Note that a more significant result (i.e., closer to zero) is favorable for corr, indicating good agreement between observations and models, but unfavorable for diff, indicating poor agreement.

	1960–99				1979–99			
	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2
	diff	diff	corr	corr	diff	diff	corr	corr
NHX	0/66	0/38	3/0	8/8	38/66	−99/−99	1/0	19/8
Tropics	0/3	0/0	−0/0	−0/−11	0/0	0/0	−0/−38	−0/−0
SHX	−19/69	−3/−0	0/0	0/0	0/0	0/0	1/0	−38/−99

tercomparison (PCMDI) for collecting and archiving the CMIP3 model output, and the WCRP's Working Group on Coupled Modelling (WGCM) for organizing the model data analysis activity. The WCRP CMIP3 multimodel dataset is supported by the Office of Science, U.S. Department of Energy. Thanks to internal reviewers Dian Seidel, Keith Dixon, Gabriel Lau, and Don Bullard, as well as journal reviewers Simon Tett and an anonymous person for their helpful comments. This research was supported in part by NOAA's Office of Global Programs.

APPENDIX

Statistical Significance

Statistical significance tests are applied to the metrics that form the basis of the bivariate plots. The philosophy adopted here is to test trends from the observations in comparison with trends from the models as a *collection*, rather than with *individual* model trends, using two different types of assessments of significance. The first type tests (a) whether the layer mean trends differ between a particular observed dataset and the model ensemble, and (b) whether there is a significant agreement between the shapes of the trend profile for a particular observed dataset and those of the model ensemble. Each assessment utilizes 21 metrics, each comparing the same observed dataset with a different model ensemble member. The following two distinct nonparametric significance tests, which are robust to outliers and non-Gaussian behavior, are performed in each case: the Wilcoxon–Mann–Whitney test (Lanzante 1996), a rank-based alternative to the student's t test, and the binomial test (Siegel and Castellan 1988). For the latter we count "hits" and "misses" when the metric of agreement is less than or greater than zero. Because the former test is more powerful, but has more restrictive assumptions, and visa versa for the latter, as a conservative approach, we report the significance level of the less significant of the two. Note that for tests involving the shape agreement, the correlation metric was transformed to the small-sample bias-corrected Fisher-z statistic (Zar 1996).

The second type of assessment aims to determine whether the observation–model agreement differs between the two different observed datasets. It is used to test whether the agreement differs between the unadjusted and adjusted versions of a given dataset, or between the adjusted versions of RATPAC and HadAT2. Again, we employ both the Wilcoxon–Mann–Whitney and binomial tests for shape agreement. However, for technical reasons the former is not applicable for testing relative bias, so we substitute a similar test, the robust rank-order test (Lanzante 1996).

Results for the first type of significance test are given in Tables A1–A2 for the troposphere and stratosphere, respectively. For the troposphere the majority of values are positive and highly significant, indicating that while the shapes of observed and model profiles are usually well correlated, their layer mean trends tend to be different, with more warming in the models. While Figs. 2 and 4 show that adjustment usually reduces the difference in layer mean trend, Table A1 indicates that in most cases this is not enough to render the disparity insignificant, a notable exception being the NHX for 1960–99. While in the extratropics the shape metric is mostly positive and often significant, in the tropics the correlations are mostly negative and usually significantly so, indicating fundamental differences between models and observations. Model profiles in the tropics (Fig. 2) are characterized by an increase in warming

TABLE A2. Same as Table A1, but for trend profiles for the stratosphere (100–30 hPa) with an additional radiosonde dataset, RW, from Randel and Wu (2006).

	1960–99						1979–99					
	RATPAC	HadAT2	RW	RATPAC	HadAT2	RW	RATPAC	HadAT2	RW	RATPAC	HadAT2	RW
	diff	diff	diff	corr	corr	corr	diff	diff	diff	corr	corr	corr
NHX	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	1/7	0/0
Tropics	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
SHX	0/0	0/0	0/0	32/79	0/1	8/99	0/0	0/0	0/0	19/19	3/−72	72/−19

TABLE A3. Similar to Table A1, but significance levels for tests of the difference between two metrics of observation–model agreement. The corr metric is as described in Table 1; however, the difference metric (adif) is a distance measure defined as the absolute value of the difference in layer mean trend, GCM minus observed. The number in each cell corresponds to a test of the difference between unadjusted and adjusted versions of the same observed dataset. The algebraic sign indicates whether the distance (adif) or correlation (corr) agreement increases (positive) or decreases (negative) in going from the unadjusted to adjusted version of the observed dataset. Note that better agreement implies a smaller difference but a larger correlation.

	1960–99				1979–99			
	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2
	adif	adif	corr	corr	adif	adif	corr	corr
NHX	9	44	5	−0	91	44	3	0
Tropics	3	3	0	0	8	3	0	0
SHX	79	−0	0	38	68	0	0	8

TABLE A4. Same as Table A3, but for the stratosphere.

	1960–99				1979–99			
	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2	RATPAC	HadAT2
	adif	adif	corr	corr	adif	adif	corr	corr
NHX	0	29	54	99	66	−0	38	99
Tropics	0	0	38	38	0	0	0	38
SHX	−99	0	−0	88	8	0	1	88

TABLE A5. Similar to Table A3, except that the number in each cell corresponds to a test of the difference between two adjusted versions of different observed datasets. The algebraic sign indicates whether the distance (adif) or correlation (corr) agreement increases (positive) or decreases (negative) in going from the first to second listed observed dataset, HadAT2 (Had) or RATPAC (RAT). Note that better agreement implies a smaller difference but a larger correlation.

	1960–99		1979–99	
	Had/RAT adif	Had/RAT corr	Had/RAT adif	Had/RAT corr
NHX	66	1	−99	0
Tropics	7	0	−74	0
SHX	0	0	66	3

TABLE A6. Same as Table A5, but for the stratosphere and with one additional observed dataset, RW (Randel and Wu 2006).

	1960–99				1979–99			
	Had/RAT	RAT/RW	Had/RAT	RAT/RW	Had/RAT	RAT/RW	Had/RAT	RAT/RW
	adif	adif	corr	corr	adif	adif	corr	corr
NHX	0	99	−2	1	−99	1	99	−38
Tropics	66	9	38	−99	−1	0	−38	−8
SHX	−99	−0	−0	79	−43	1	66	−3

Table A7. A mapping of the correspondence between Table 1 and Tables A1–A6. There is a one-to-one correspondence between the first two columns here (troposphere, stratosphere), and similar columns in Table 1, as well as the six rows in each table. In each cell (for the first two columns in this table) the first element indicates the appendix table number while the second indicates the metric (diff/corr/adiff). Each of the entries in Table 1 represents an aggregate over a number of sets. The column labeled "N" here indicates the total number of items (significance test results) in the aggregation while that labeled "Sets" gives the specific sets used in the aggregations. The sets include the time periods for the trends (1960–99/1979–99), the latitude zones (NHX/tropics/SHX), the datasets (RATPAC/HadAT2), and the homogeneity treatments (unadjusted/adjusted).

Troposphere	Stratosphere	N	Sets
Table A1 diff	Table A2 diff	24	Time period, zone, dataset, treatment
Table A1 corr	Table A2 corr	24	Time period, zone, dataset, treatment
Table A3 adiff	Table A4 adiff	12	Time period, zone, dataset
Table A3 corr	Table A4 corr	12	Time period, zone, dataset
Table A5 diff	Table A6 diff	6	Time period, zone
Table A5 corr	Table A6 corr	6	Time period, zone

from the lower to upper troposphere with warming throughout the upper troposphere. In contrast, observed profiles have tropospheric maxima at lower levels than the models, and cool in the upper troposphere.

Significance tests for the stratosphere (Table A2) are quite one sided in that most cases are highly significant for both unadjusted and adjusted versions. As for the troposphere the relative bias is positive, however here the shape agreement is much better. The significances of the differences and correlations for the RW (reduced) network of stations are in most cases quite similar to those from the full RATPAC network.

Results of tests of differences in observation–model agreement between unadjusted and adjusted versions of the same dataset are presented in Tables A3–A4. Note that in contrast to earlier analyses, the difference metric used here is a measure of distance based on the absolute value of the observation–model difference. By convention, a positive (negative) value in these tables indicates better (poorer) agreement, in terms of distance or shape, in going from the unadjusted to adjusted version of the dataset. The values in Table A3 for the troposphere are striking in that almost all are positive, indicating improved agreement with adjustment. The majority of values are significant as well. Note that the only degradation with adjustment occurs for HadAT2 for separate instances in the extratropics for 1960–99. Adjustment enhances agreement more with regards to shape than distance, and more so in the tropics than in the extratropics. Results for the stratosphere (Table A4) also indicate that adjustment has an overwhelming tendency to enhance agreement, again, especially in the tropics. However, there is a reversal in that the enhancement via adjustment is much more prominent with regard to the distance metric than with regards to shape.

The final sets of significance assessments test for differences in observation–model agreement between adjusted versions of the two observed datasets. The convention used in Table A5 for the troposphere is such that positive values indicate better agreement for RATPAC than HadAT2. With regard to shape agreement the results are quite one sided, such that RATPAC yields significantly better agreement than HadAT2 in every case. There is much less distinction between the two observed datasets with regard to the distance metric, although the two most significant results also indicate better agreement for RATPAC. Results for the stratosphere (Table A6) are less one sided. For the distance metric most results are not significant and there is no obvious preference for either dataset. For shape the signs are equally split and most results are not significant, although the two most significant results are more favorable for HadAT2. Thus, overall, in the stratosphere HadAT2 displays slightly better agreement with the models.

Table A6 also reports comparisons between the full RATPAC network and the reduced RW network, with positive values indicating an improvement for RW. The RW network yields overwhelmingly better agreement with regards to distance, especially during the satellite era. For the shape metric RW yields poorer agreement, particularly during the satellite era. Note that because Randel and Wu (2006) used deep layer mean temperatures in their assessments, it is not surprising that RW has more favorable agreement with regard to the distance metric, because they could not resolve vertical structure. It is also not surprising that the better distance agreement occurs during the satellite era, the period over which they performed their analyses.

Table A7 indicates how the summary in Table 1 was constructed from the detailed results in Tables A1–A6. For example, the first row and column of Table A7 indicate that difference metrics (diff) from Table A1 were used to derive the 67% value indicated in the first row of Table 1 for the troposphere. Taking 67% of the 24 total items (N from first row of Table A7) yields 16 as the number of tests in the aggregation found to be significant at the 5% level. The right-most column of Table A7 indicates the significance test cases included in the aggregation. The interested reader can use the reminder of Table A7 as a guide in constructing the entirety of Table 1.

REFERENCES

Brohan, P., J. J. Kennedy, I. Haris, S. F. B. Tett and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Cordero, E. C., and P. M. de F. Forster, 2006: Stratospheric variability and trends in models used for the IPCC AR4. *Atmos. Chem. Phys.,* **6,** 5369–5380.

Folland, C. K., D. M. H. Sexton, D. J. Karoly, C. E. Johnson, D. P. Rowell, and D. E. Parker, 1998: Influences of anthropogenic and oceanic forcing on recent climate change. *Geophys. Res. Lett.,* **25,** 353–356.

Free, M., J. K. Angell, I. Durre, J. R. Lanzante, T. C. Peterson, and D. J. Seidel, 2004: Using first differences to reduce inhomogeneity in radiosonde temperature datasets. *J. Climate,* **17,** 4171–4179.

——, D. J. Seidel, J. K. Angell, J. Lanzante, I. Durre, and T. C. Peterson, 2005: Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new dataset of large-area anomaly time series. *J. Geophys. Res.,* **110,** D22101, doi:10.1029/2005JD006169.

Gaffen, D. J., 1994: Temporal inhomogeneities in radiosonde temperature records. *J. Geophys. Res.,* **99,** 3667–3676.

Hansen, J., and Coauthors, 2002: Climate forcings in Goddard Institute for Space Studies SI2000 simulations. *J. Geophys. Res.,* **107,** 4347, doi:10.1029/2001JD001143.

Joshi, M. M., and K. P. Shine, 2003: A GCM study of volcanic eruptions as a cause of increased stratospheric water vapor. *J. Climate,* **16,** 3525–3534.

Karl, T. R., S. J. Hassol, C. D. Miller, and W. L. Murray, Eds., 2006: Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences. U.S. Climate Change Science Program and the Subcommittee on Global Change Research Final Report, Synthesis and Assessment Product 1.1, 180 pp. [Available online at http://www.climatescience.gov/Library/sap/sap1-1/finalreport/default.htm.]

Karoly, D. J., 1987: Southern Hemisphere temperature trends: A possible greenhouse gas effect? *Geophys. Res. Lett.,* **14,** 1139–1141.

——, 1989: Northern Hemisphere temperature trends: A possible greenhouse gas effect? *Geophys. Res. Lett.,* **16,** 465–468.

Lanzante, J. R., 1996: Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.,* **16,** 1197–1226.

——, 2007: Diagnosis of radiosonde vertical temperature trend profiles: Comparing the influence of data homogenization versus model forcings. *J. Climate,* **20,** 5356–5364.

——, S. A. Klein, and D. J. Seidel, 2003a: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

——, ——, and ——, 2003b: Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. *J. Climate,* **16,** 241–262.

Madden, R. A., and V. Ramanathan, 1980: Detecting climate change due to increasing carbon dioxide. *Science,* **209,** 763–768.

McCarthy, M. P., H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2008: Assessing bias and uncertainty in the HadAT-adjusted radiosonde climate record. *J. Climate,* **21,** 817–832.

National Research Council, 2000: *Reconciling Observations of Global Temperature Change*. National Academy Press, 85 pp.

Parker, D. E., M. Gordon, D. P. N. Cullum, D. M. H. Sexton, C. K. Folland, and N. Rayner, 1997: A new global gridded radiosonde temperature database and recent temperature trends. *Geophys. Res. Lett.,* **24,** 1499–1502.

Ramaswamy, V., M. D. Schwarzkopf, W. J. Randel, B. D. Santer, B. J. Soden, and G. L. Stenchikov, 2006: Anthropogenic and natural influences in the evolution of lower stratospheric cooling. *Science,* **311,** 1138–1141.

Randel, W. J., and F. Wu, 2006: Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *J. Climate,* **19,** 2094–2104.

Santer, B. D., and Coauthors, 1996: A search for human influences on the thermal structure of the atmosphere. *Nature,* **382,** 39–46.

——, and Coauthors, 2001: Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. *J. Geophys. Res.,* **106** (D22), 28 033–28 059.

——, and Coauthors, 2003: Influence of satellite data uncertainties on the detection of externally forced climate change. *Science,* **300,** 1280–1284.

——, and Coauthors, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science,* **309,** 1551–1556.

Seidel, D. J. and J. R. Lanzante, 2004: An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *J. Geophys. Res.,* **109,** D14108, doi:10.1029/2003JD004414.

Sexton, D. M. H., D. P. Rowell, C. K. Folland, and D. J. Karoly, 2001: Detection of anthropogenic climate change using an atmospheric GCM. *Climate Dyn.,* **17,** 669–685.

Sherwood, S. C., J. R. Lanzante, and C. L. Meyer, 2005: Radiosonde daytime biases and late-20th Century warming. *Science,* **309,** 1556–1559.

——, C. L. Meyer, R. J. Allen, and H. A. Titchner, 2008: Robust tropospheric warming revealed by iteratively homogenized radiosonde data. *J. Climate,* **21,** 5336–5352.

Shine, K. P., and Coauthors, 2003: A comparison of model-simulated trends in stratospheric temperatures. *Quart. J. Roy. Meteor. Soc.,* **129,** 1565–1588.

Siegel, S., and N. Castellan, 1988: *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, 399 pp.

Solomon, S., R. W. Portmann, R. R. Garcia, L. W. Thomason, L. R. Poole, and M. P. McCormick, 1996: The role of aerosol variations in anthropogenic ozone depletion at northern midlatitudes. *J. Geophys. Res.,* **101** (D3), 6713–6728.

Stott, P. A., and S. F. B. Tett, 1998: Scale-dependent detection of climate change. *J. Climate,* **11,** 3282–3294.

——, G. S. Jones, J. A. Lowe, P. Thorne, C. Durman, T. C. Johns, and J.-C. Thelen, 2006: Transient climate simulations with the HadGEM1 climate model: Causes of past warming and future climate change. *J. Climate,* **19,** 2763–2782.

Tett, S. F. B., J. F. B. Mitchell, D. E. Parker, and M. R. Allen, 1996: Human influence on the atmospheric vertical temperature structure: Detection and observations. *Science,* **274,** 1170–1173.

——, and Coauthors, 2002: Estimation of natural and anthropogenic contributions to twentieth century temperature change. *J. Geophys. Res.,* **107,** 4306, doi:10.1029/2000JD000028.

Thorne, P. W., and Coauthors, 2003: Probable causes of late twentieth century tropospheric temperature trends. *Climate Dyn.,* **21** (7–8), 573–591.

——, D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005: Revisiting radiosonde upper air temperatures from 1958 to 2002. *J. Geophys. Res.,* **110,** D18105, doi:10.1029/2004JD005753.

Wittenberg, A. T., A. Rosati, N.-C. Lau, and J. J. Ploshay, 2006: GFDL's CM2 Global Coupled Climate Models. Part III: Tropical Pacific climate and ENSO. *J. Climate,* **19,** 698–722.

WMO, 2006: Scientific assessment of ozone depletion: 2006. Global Ozone Research and Monitoring Project, Report 50, 662 pp.

Zar, J. H., 1996: *Biostatistical Analysis.* 3rd ed. Prentice Hall, 572 pp.